

# LINDAT/CLARIAH-CZ

PAVEL STRAŇÁK

Repository

[lindat.cz/repository](http://lindat.cz/repository)



# LINDAT/CLARIAH-CZ


PAVEL STRAŇÁK

FAIR repository for language data  
+ humanities' datasets

[lindat.cz/repository](https://lindat.cz/repository)

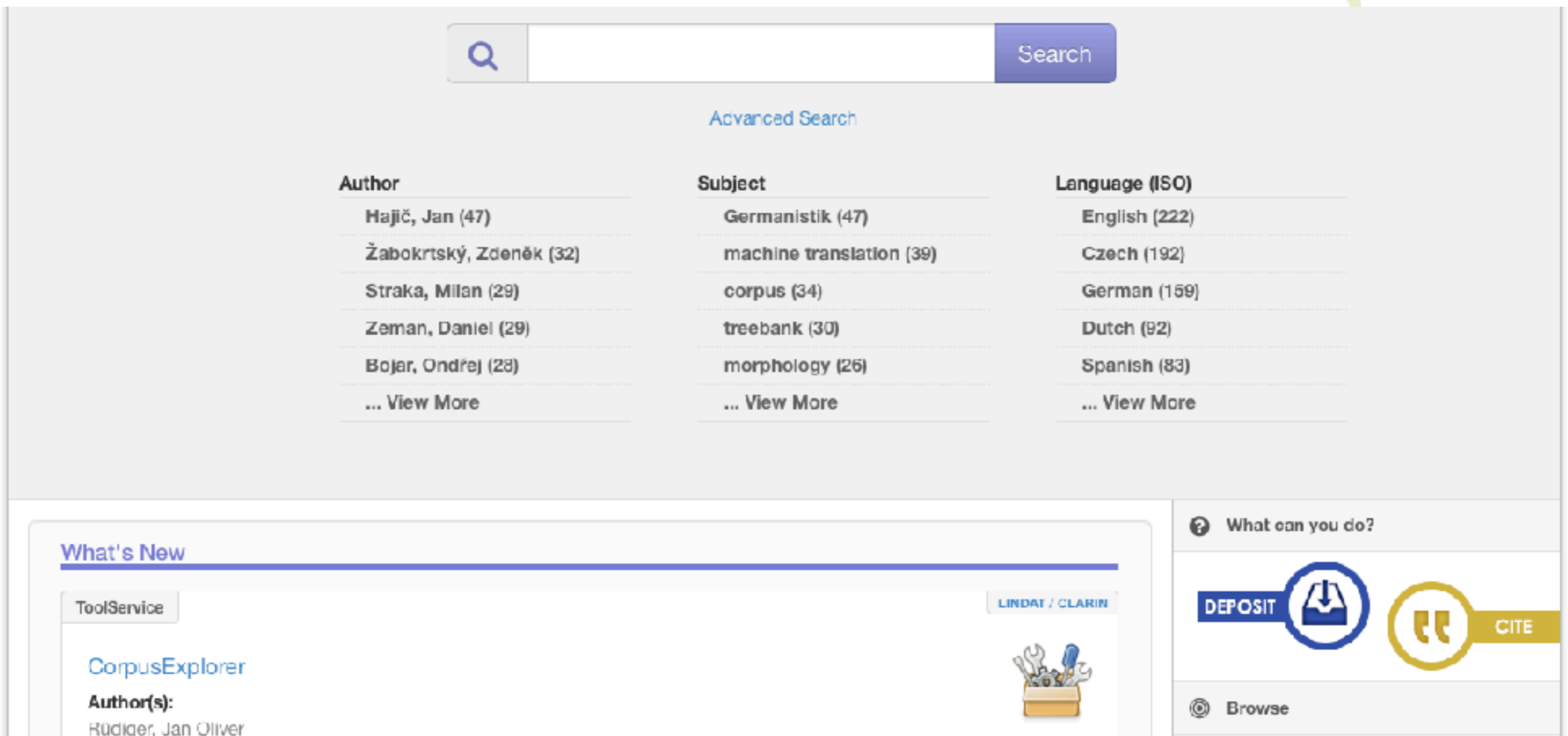


# CLARIN-DSpace GitHub project

- <https://github.com/ufal/clarin-dspace> 
- About 15 installations in 11 countries
  - Some installations just using the Github documentation and #dspace at CLARIN Slack
  - Some installations by LINDAT team using CLARIN Mobility grants. With some preparation in advance, a working B-centre compliant repository can be up and running in 2 days.
- Some forks (Polish, Norwegian) have a few additional features
- Most centres probably motivated by taking a ready-made repository
  - The collaborative development (naturally) works a bit less great than we had hoped
  - Upcoming upgrade to Dspace 7 (new UI) will be the real test of the community

# Data Repository

Preserve and find language data and NLP tools



The screenshot displays the Data Repository website interface. At the top, there is a search bar with a magnifying glass icon and a "Search" button. Below the search bar is a link for "Advanced Search". The main content area is divided into three columns: "Author", "Subject", and "Language (ISO)". Each column lists several items with their respective counts in parentheses. The "Author" column lists Hajič, Jan (47), Žabokrtský, Zdeněk (32), Straka, Milan (29), Zeman, Daniel (29), and Bojar, Ondřej (28), with a "View More" link. The "Subject" column lists Germanistik (47), machine translation (39), corpus (34), treebank (30), and morphology (26), with a "View More" link. The "Language (ISO)" column lists English (222), Czech (192), German (159), Dutch (92), and Spanish (83), with a "View More" link. Below the main content area, there is a "What's New" section with a "ToolService" tab and a "LINDAT / CLARIN" logo. The "What's New" section features a "CorpusExplorer" tool by Rüdiger, Jan Oliver, accompanied by an icon of a toolbox. To the right of the "What's New" section, there is a "What can you do?" section with icons for "DEPOSIT" (a blue circle with a white icon of a document being placed into a folder), "CITE" (a yellow circle with a white icon of quotation marks), and "Browse" (a grey circle with a white icon of a magnifying glass).

Author	Subject	Language (ISO)
Hajič, Jan (47)	Germanistik (47)	English (222)
Žabokrtský, Zdeněk (32)	machine translation (39)	Czech (192)
Straka, Milan (29)	corpus (34)	German (159)
Zeman, Daniel (29)	treebank (30)	Dutch (92)
Bojar, Ondřej (28)	morphology (26)	Spanish (83)
... View More	... View More	... View More



**What's New**


ToolService LINDAT / CLARIN

**CorpusExplorer**


**Author(s):**  
Rüdiger, Jan Oliver

**What can you do?**

**DEPOSIT**  **CITE** 

**Browse** 

# Data Repository

- **OPEN**  **ACCESS** (as can be – [Public License Selector](#))
- > 500 registered users
  - submitters & users signing licenses (not everything can be OA)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

# Data Repository

- > 500 registered users
  - submitters & users signing licenses (not everything can be Open Access)
- 200+ Data Records
  - > 1000 Metadata Records
  - 80 languages
- 100 TB+ Data in Repository (+ 1PB of UCS Shoah Foundation Archive)

The screenshot shows the LINDAT/CLARIN Repository Home page. At the top, there is a navigation bar with the LINDAT logo and links for Repository, TreeQuery, Treex, More Apps, and About. Below the navigation bar is a search bar with a magnifying glass icon and a 'Search' button. Underneath the search bar is an 'Advanced Search' link. The main content area is divided into two columns. The left column is titled 'Limit your search' and contains several dropdown menus: Author, Subject, Rights, Language (ISO), Type, Contain Files, and Community. The right column is titled 'Showing 1 through 10 out of 1038 results' and contains a list of search results. The first result is 'AKCES 2 ver. 2' (Charles University in Prague, ÚČJTK / 2013-12-18) by Šebesta, Karel ; Golářová, Hana. It is publicly available and contains 1 file (3.85 MB). The second result is 'A Gold Standard Word Alignment for English-Swedish (2015-10-12)' (Linköping University / 2015-10-12) by Ahrenberg, Lars ; Holmqvist, Maria. It is publicly available and contains 1 file (590 KB). The third result is 'MorphoDiTa: Morphological Dictionary and Tagger' (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14) by Straka, Milan ; Straková, Jana. It contains no files.

# Data Repository

- Safe preservation (upload and don't worry)
- Discovery & Reuse
- Direct data citation (works in Google Scholar)
- Licensing (Open Access, but also more options)
- Versioning
- Worldwide (for everyone), easy to use

The screenshot shows the LINDAT/CLARIN Repository website. At the top, there is a navigation bar with the LINDAT logo and links for Repository, TreeQuery, Treex, More Apps, and About. Below the navigation bar is a search bar with a magnifying glass icon and a 'Search' button. Underneath the search bar is a link for 'Advanced Search'. The main content area is divided into two columns. The left column is titled 'Limit your search' and contains several dropdown menus for filtering results: Author, Subject, Rights, Language (ISO), Type, Contain Files, and Community. The right column is titled 'Showing 1 through 10 out of 1038 results' and displays a list of search results. The first result is 'AKCES 2 ver. 2' by Charles University in Prague, with authors Šebesta, Karel and Golářová, Hana. The second result is 'A Gold Standard Word Alignment for English-Swedish (2015-10-12)' by Linköping University, with authors Ahrenberg, Lars and Holmquist, Maria. The third result is 'MorphoDiTa: Morphological Dictionary and Tagger' by Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL), with authors Straka, Milan and Straková, Jana. Each result includes a 'Publicly Available' badge and a link to the item's page.

# Safe Preservation

## Upload and don't worry

### How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at [clarin.eu](http://clarin.eu) and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

### Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.

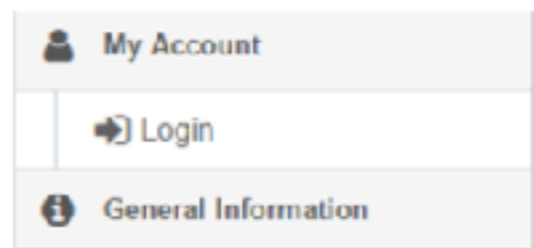


Fig1. Menu Login

### Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.

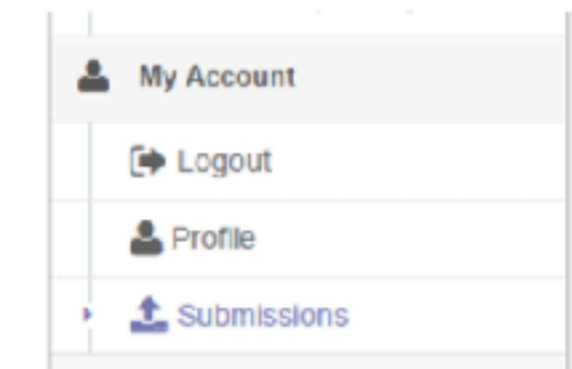
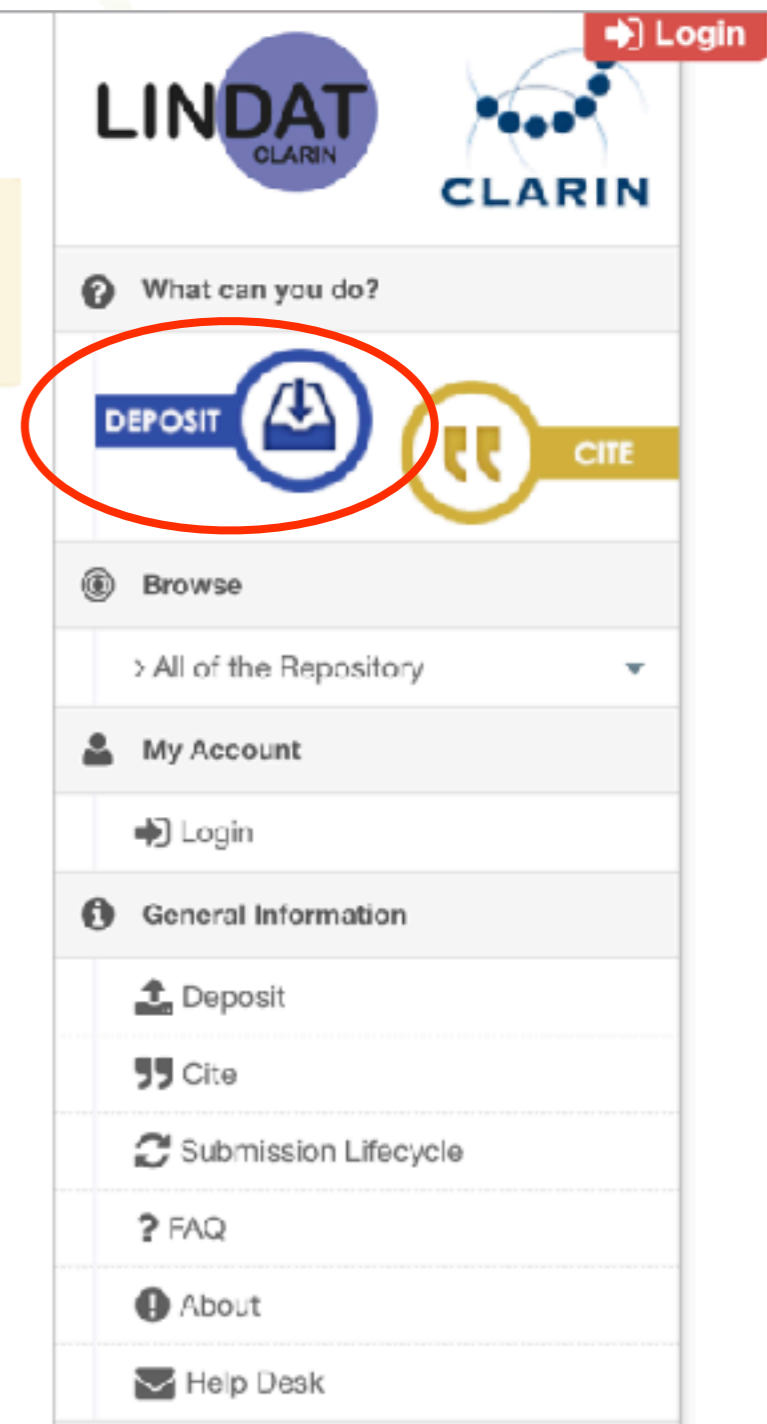


Fig2. Menu Submissions





# Deposition Guide

## step-by-step description

1. login
2. fill-in metadata
3. drag&drop data
4. select a license
5. submit

## Step 3: Select type of your submission

You have initiated a new workflow item. In the next few steps you will provide the details about the item, select the type of the resource you are about to submit.

Item submission

1. Basic info 2. Who's involved 3. Describe 4. Upload 5. License 6. Note 7. Review

Submission Info

Corpus Lexical conceptual Language description Technology / Tool / Service

Type of the resource: "Corpus" refers to text, speech and multimodal corpora. "Lexical Conceptual Resource" includes lexica, ontologies, dictionaries, word lists etc. "Language Description" refers to models and grammars. "Technology / Tool / Service" is used for tools, systems, system components.

Title

Enter the main title of the item in English.

Fig4. Submission info

Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as title in the following step.

## Step 4: Describe your item

In the following two steps you will provide more details for your item. First describe the people, organization, the item.

Item submission

1. Basic info 2. Who's involved 3. Describe 4. Upload 5. License 6. Note 7. Review

# Login to Deposit


- institutional logins (EduID-cz, EduGAIN)
- CLARIN account for the “homeless researchers”
- minimal personal info


Sign in to **LINDAT/CLARIN Repository**  
Login via Your home institution (e.g. university)

- Univerzita Karlova v Praze  
Czech Republic 6 km
- Institute of Biotechnology CAS, v.v.i.  
Czech Republic 4 km
- Czech University of Life Sciences Prague  
Czech Republic 4 km
- Institute of Art History of the Academy of Sciences of the Czech Republic  
Czech Republic 4 km
- College of Polytechnics Jihlava  
Czech Republic 4 km
- Global Change Research Institute CAS  
Czech Republic 4 km
- Czech Language Institute of the Czech Academy of Sciences  
Czech Republic 4 km
- Institute of Chemical Process Fundamentals of the AS CR  
Czech Republic 4 km
- Institute of Theoretical and Applied Mechanics AS CR

Q:pr search for a provider, such as Example University

▶ **Please help, I cannot find my provider**

 Locate me and show nearby providers

Show providers in   show all countries

# Faceted Search

Advanced Search

Search

Limit your search

- Author
- Subject
- Rights
- Language (ISO)
- Type
- Contain Files
- Community

Showing 1 through 10 out of 1038 results

1 2 3 > 104

Corpus LINDAT / CLARIN

**AKCES 2 ver. 2**  
(Charles University in Prague, ÚČJTK / 2013-12-18)

**Author(s):**  
Šebesta, Karel ; Goláňová, Hana

This item contains 1 file (3.85 MB).

Publicly Available

LexicalConceptualResource LRT + Open Submissions

**A Gold Standard Word Alignment for English-Swedish (2015-10-12)**  
(Linköping University / 2015-10-12)

**Author(s):**  
Ahrenberg, Lars ; Holmqvist, Maria

This item contains 1 file (590 KB).

LINDAT CLARIN

What can you do?

DEPOSIT CITE

Browse

> All of the Repository

My Account

Login

General Information

- Deposit
- Cite
- Submission Lifecycle
- FAQ
- About
- Help Desk

# Faceted Search

  [Search](#)

[Advanced Search](#)

**Limit your search**

- [Author](#)
- [Subject](#)
- [Rights](#)
- [Language \(ISO\)](#)
- [Type](#)
- [Contain Files](#)
- [Community](#)

Showing 1 through 10 out of 1038 results

1 2 3 > 104

**Corpus** LINDAT / CLARIN

[AKCES 2 ver. 2](#) 


(Charles University in Prague, ÚČJTK / 2013-12-18)

**Author(s):**  
Šebesta, Karel ; Goláňová, Hana

[This item contains 1 file \(3.85 MB\).](#)

**Publicly Available** 

**LexicalConceptualResource** LRT + Open Submissions



[A Gold Standard Word Alignment for English-Swedish \(2015-10-12\)](#) 

(Linköping University / 2015-10-12)

**Author(s):**  
Ahrenberg, Lars ; Holmqvist, Maria

[This item contains 1 file \(590 KB\).](#)



- [What can you do?](#)
- [DEPOSIT](#)  [CITE](#) 
- [Browse](#)
- > All of the Repository
- [My Account](#)
- [Login](#)
- [General Information](#)
- [Deposit](#)
  - [Cite](#)
  - [Submission Lifecycle](#)
  - [FAQ](#)
  - [About](#)
  - [Help Desk](#)

# Discovery

Google



prague dependency treebank 3.0



Vše

Obrázky

Nákupy

Mapy

Videa

Více

Nastavení

Nástroje

Přibližný počet výsledků: 13 900 (0,44 s)

## Vědecké články o prague dependency treebank 3.0

**Prague dependency treebank 3.0** - **Bejček** - Počet citací tohoto článku: 47

**Prague Dependency Treebank** - **Hajič** - Počet citací tohoto článku: 385

**The Prague dependency treebank** - **Böhmová** - Počet citací tohoto článku: 423

## Prague Dependency Treebank 3.0 | ÚFAL

<https://ufal.mff.cuni.cz/pdt3.0> ▼ Přeložit tuto stránku

Introduction. The Prague Dependency Treebank 3.0 (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed and improved in various aspects. Moreover ...

## The Prague Dependency Treebank 2.0.

<https://ufal.mff.cuni.cz/pdt2.0/> ▼ Přeložit tuto stránku

The Prague Dependency Treebank 2.0 (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation ... Please note that new versions of this corpus have been published: PDT 3.0 (2013), PDiT 1.0 (2012), PDT 2.5 (2012).

## Prague Dependency Treebank 3.0 (PDT 3.0)

[https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30\\_index\\_lindat.html?...](https://lindat.mff.cuni.cz/repository/xmlui/bitstream/.../PDT30_index_lindat.html?...)

Prague Dependency Treebank 3.0 (PDT 3.0). Overview. The Prague Dependency Treebank 3.0 (PDT 3.0) annotates the same texts as the PDT 2.0 (Hajič et al. 2006), PDT 2.5 (Bejček et al. 2011), and the Prague Discourse Treebank 1.0 (PDiT 1.0, Poláková et al. 2012). The annotation on the four layers was further fixed ...

# Direct Data Citations

## Credit for Data

enTenTen

Please use the following text to cite this item or export to a predefined format:

BIBTEX CMDI

Masaryk University, NLP Centre, 2011, *enTenTen*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8>.

This resource is also integrated in following services:

Share:

KonText

Item identifier	<a href="http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8">http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8</a>
Date issued	2011-12-16
Type	corpus
Language(s)	English
Description	Very large English web corpus enTenTen, comprising 3,268,798,627 tokens.
Publisher	Masaryk University, NLP Centre
Acknowledgement	Lexical Computing Ltd.
Subject(s)	English large corpus
Collection(s)	LINDAT / CLARIN Data & Tools

[Show full item record](#)

LINDAT  
CLARIN

Pavel Straňák | [Logout](#)

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Context

> Edit this item

> Export Item

> Export Metadata

Administrative

Control Panel

Access Control

# Licensing

## As Open as Possible

### Choose a License

Answer the questions or use the search to find the license you want

[Start again](#) ← →

What do you want to deposit?


[Software](#) [Data](#)

Search for a license...

---

**Public Domain Mark (PD)**



The work identified as being free of known restrictions under copyright law, including all related and neighboring rights.

[Publicly Available](#) 

---

**Public Domain Dedication (CC Zero)**



CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

[Publicly Available](#)   [OPEN DATA](#)

---

**Creative Commons Attribution (CC-BY)**

This is the standard creative commons license that gives others maximum freedom to do what they want with your work.

[Publicly Available](#)   [OPEN DATA](#)

# Licensing

As Open as Possible (not more)

 **Publisher** Faculty of Arts, Institute of the Czech National Corpus, Charles University in Prague

 **Acknowledgement** Ministerstvo školství, mládeže a tělovýchovy

Project code: LM2011023

Project name: Český národní korpus

 **Subject(s)** representative corpus written language

 **Collection(s)** LINDAT / GLARIN Data & Tools

[Show full item record](#)

 **Files in this item**



Download instructions for command line

This item is **Academic Use** and licensed under:  
Czech National Corpus (Shuffled Corpus Data)



<b>Name</b>	syn2015.gz
<b>Size</b>	1.48 GB
<b>Format</b>	application/x-gzip
<b>Description</b>	corpus
<b>MD5</b>	e0242cc77e999794af6cfaf57f843c12



 [Download file](#)

CLEAR RULES  
CUSTOM LICENSES  
LICENSE SIGNING



# Licensing Framework:

Any License (Open Source / Open Data preferred)



## Manage Licenses

All Licenses
Define License
Define License Label

	License Name	Definition (URL)	Confirma- tion	Required user info	Licen- se Labe- l	Exte- nded Labe- ls	Used by Bitstr- eams
<input type="checkbox"/>	Universal Derivations v0.5 License Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-0.5">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-0.5</a>	Not required		PUB	CC	1
<input type="checkbox"/>	Licence Universal Dependencies v2.4	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.4">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.4</a>	Not required		PUB	CC GPLv3	4
<input type="checkbox"/>	License agreement for The Multilingual corpus of literal occurrences of multiword expressions	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-literal">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mwe-literal</a>	Not required		PUB	CC GPLv3	5
<input type="checkbox"/>	Licence Universal Dependencies v2.3	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.3">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.3</a>	Not required		PUB	CC GPLv3 GPLv2	4
<input type="checkbox"/>	PARSEME Shared Task Data (v. 1.1) Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mws-1.1">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mws-1.1</a>	Not required		PUB	CC GPLv3	22
<input type="checkbox"/>	Licence Universal Dependencies v2.2	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.2">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.2</a>	Not required		PUB	CC GPLv3 GPLv2	7
<input type="checkbox"/>	Licence Universal Dependencies v2.1	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.1">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.1</a>	Not required		PUB	CC GPLv3 GPLv2	4
<input type="checkbox"/>	PARSEME Shared Task Data (v. 1.0) Agreement	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mws-1.0">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-mws-1.0</a>	Not required		PUB	CC GPLv3	21
<input type="checkbox"/>	Licence Universal Dependencies v2.0	<a href="https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.0">https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.0</a>	Not required		PUB	CC GPLv3 GPLv2	14

Pavel Stranák | Logout

What can you do?

DEPOSIT

CITE

Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

Registries

Collections & Communities

Statistics

Curation Tasks

# Licensing Framework

## Defining a New License



### Manage Licenses

[All Licenses](#)[Define License](#)[Define License Label](#)

#### Define new licence

License name

License definition URL

License requires confirmation

License Labels

- Publicly Available (PUB)  
 Academic Use (ACA)  
 Restricted Use (RES)

License Labels

- Attribution Required (BY)  
 Share Alike (SA)  
 Noncommercial (NC)  
 Redeposit Modified (ReD)  
 No Derivative Works (ND)  
 Inform Before Use (Inf)  
 Distributed under Creative Commons (CC)  
 No Copyright (ZERO)  
 GNU General Public License, version 3.0 (GPLv3)  
 GNU General Public License, version 2.0 (GPLv2)  
 BSD (BSD)  
 The MIT License (MIT)  
 The Open Source Initiative (OSI)

Additional required user info

- The user will receive an email with download instructions.  
 User Name  
 Date of Birth  
 Address  
 Country  
 Ask user for another email address

Pavel Straňák | Logout



What can you do?



Browse

&gt; All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

&gt; Content Administration

Registries

Collections &amp; Communities

Statistics

Curation Tasks

Licenses

Handles

# Licensing Framework

## Defining a licensing label / attribute

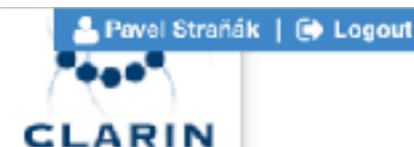


### Manage Licenses

[All Licenses](#)[Define License](#)[Define License Label](#)

#### Define New Label

Short Label	<input type="text"/>
Label Title	<input type="text"/>
Is extended	<input type="text" value="Yes"/>
Icon image	<input type="button" value="Choose File"/> no file selected
	<input type="button" value="Save"/> <input type="button" value="Return"/>



What can you do?



Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

Registries

Collections & Communities

Statistics

Curation Tasks

Licenses

Handles

# Versioning

prefer latest, preserve all

Project name: Internet jako jazykový korpus

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: LN00A063

Project name: Centrum počítační lingvistiky

Ministerstvo školství, mládeže a tělovýchovy České republiky

Project code: MSM 0021620838

Project name: Moderní metody, struktury a systémy informatiky

## Subject(s)

MorphoDiTa

Czech

morphological analysis

morphological generation

PoS tagging

## Collection(s)

LINDAT / CLARIN Data & Tools



This item is replaced by a newer submission:

<http://hdl.handle.net/11234/1-1836>

Please refer to the submission above for the latest available data. If you nevertheless need the original data, please click [here](#).

List all versions ▼

# Versioning

prefer latest, preserve all

## Other versions

List all versions ▾

- ▶ Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
- Czech Models (Morfflex CZ 160310 + PDT 3.0) for MorphoDiTa 160310
- Czech Models (Morfflex CZ + PDT) for MorphoDiTa

[Show full item record](#)

## Files in this item



Download instructions for command line

This item is **Publicly Available** and licensed under:

Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)



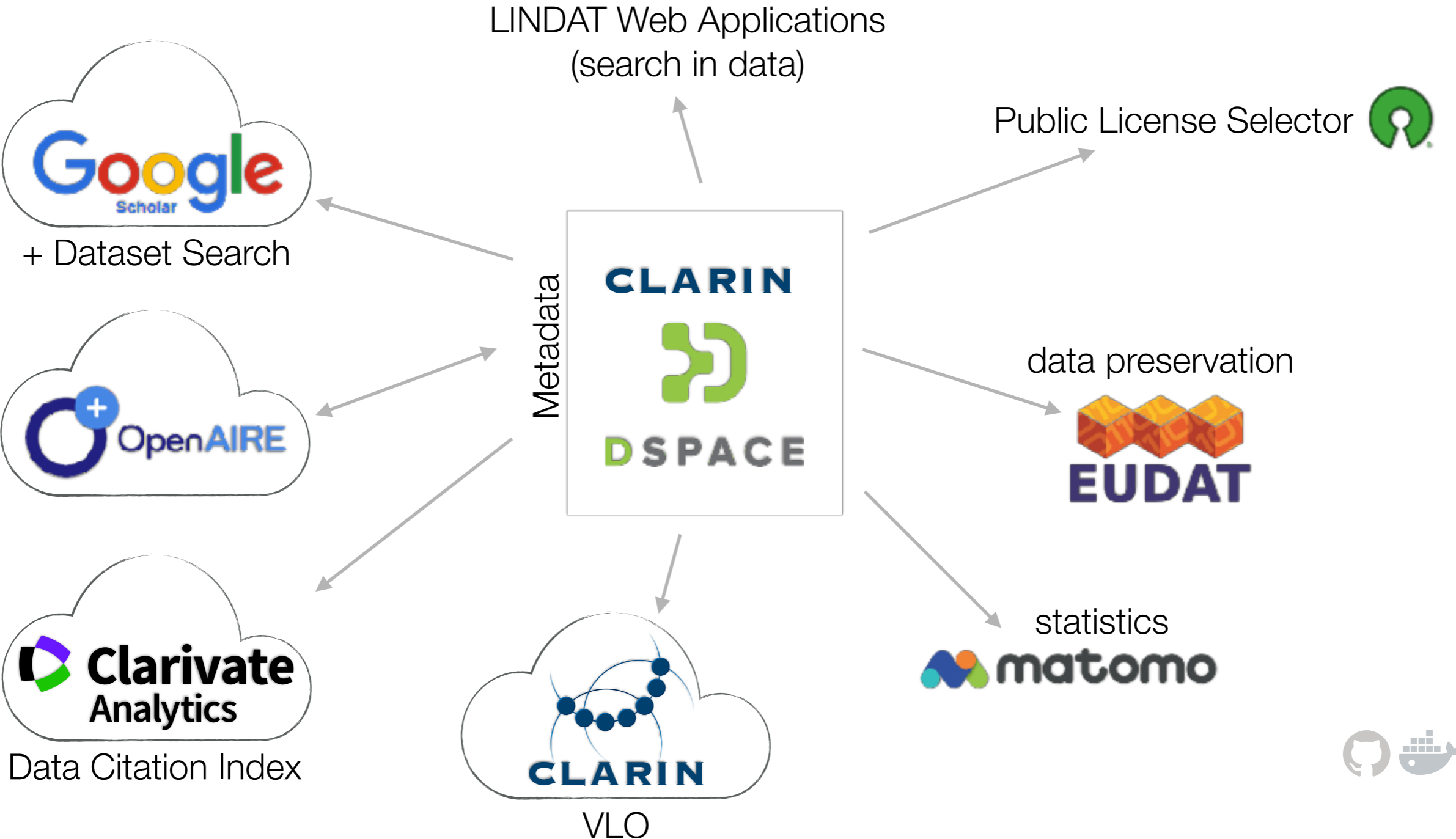
<b>Name</b>	czech-morfflex-pdt-161115.zip
<b>Size</b>	69.18 MB
<b>Format</b>	application/zip
<b>Description</b>	Czech Models (Morfflex CZ 161115 + PDT 3.0) for MorphoDiTa 161115
<b>MD5</b>	adde38cd363219759e19165b06baa4ce



[Download file](#)

[Preview](#)

# Repository Integrations



# FAIR summary

- **Findable**: Google, Google Scholar, Data Citation Index, CLARIN VLO, OLAC... and the repository itself
- **Accessible**: records with data (even when restricted), complete licensing, Open Access (Public License Selector), login only when needed, CESNET, EUDAT
- **Interoperable**: common data formats, full documentation (enhanced metadata, documentation bitstreams)
- **Reusable**: records with data, complete licensing, full versioning, direct data citations, maximal OA

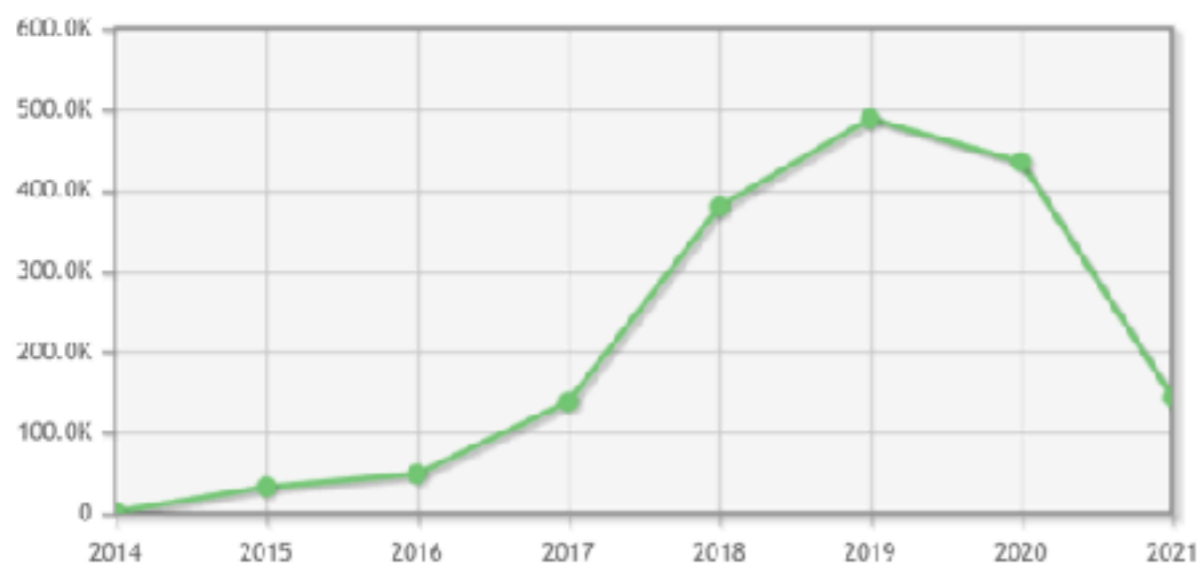
# Statistics

- 1123 records, 436 containing data:
  - 353 from LINDAT, 83 in LRT Inventory (worldwide open submissions)

The download statistics are based on the tracking request PIWIK received from the JAVA API directly from DSpace. The download counts also contain hits from bots.

## Downloads Over Time

2014 - 2021



Click on a data item to summarize by year / month.

## Important Metrics

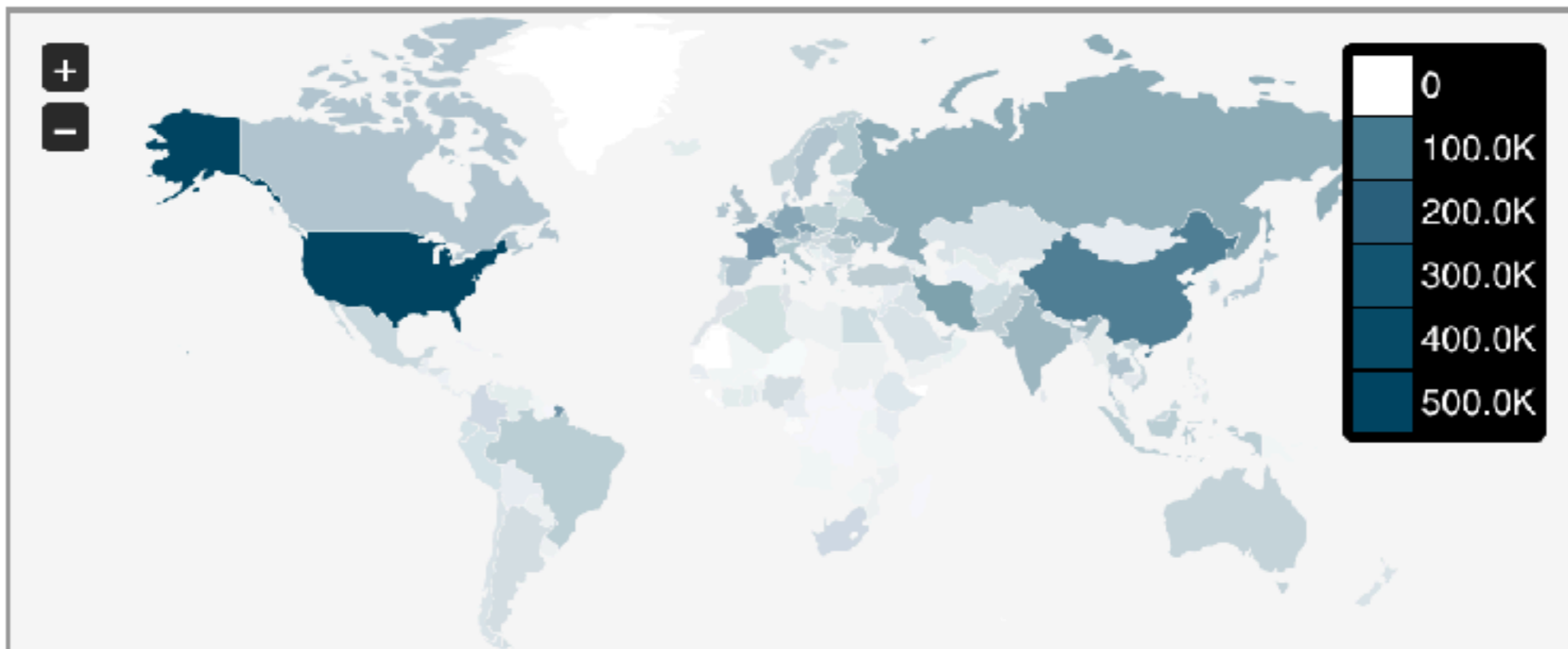
<u>Downloads</u>	<b>1665033</b>	<u>Unique Downloads</u>	<b>1115926</b>
<u>Visits</u>	<b>717535</b>	<u>Unique Visitors</u>	<b>624774</b>

Hover on a metric name to show its definition.



# Statistics

## Country Wise Visits



 Visits from Czech Republic **16620 (2.32%)**

Thank you!

<https://lindat.cz/repository>

